
Cognitive Science and Thought Experiments: A Refutation of Paul Thagard's Skepticism

Michael T. Stuart

University of Toronto

Paul Thagard has recently argued that thought experiments are dangerous and misleading when we try to use them as evidence for claims. This paper refutes his skepticism. Building on Thagard's own work in cognitive science, I suggest that Thagard has much that is positive to say about how thought experiments work. My last section presents some new directions for research on the intersection between thought experiments and cognitive science.

Paul Thagard is a well-known cognitive scientist and philosopher of mind who has recently expressed skepticism about the cognitive efficacy of thought experiments.¹ In so doing he joins forces with Alexius Meinong (1907), Daniel Dennett (1984), Jonathan Dancy (1985), Gilbert Harman (1986), and Kathleen Wilkes (1988). According to Meinong, who was perhaps the first skeptic about thought experiments explicitly so-called, “an experiment that in fact does not exist at all, can neither prove nor teach anything” (1907, pp. 276–77). Dennett, Dancy, Harman and Wilkes are no less forceful. What sets Thagard apart is the *source* of his skepticism: cognitive science. This is noteworthy because cognitive science is heavily drawn upon by the largest and fastest growing group of people writing about thought experiments today, see, e.g., Bishop (1998), Cooper (2005), Gendler (2004), Palmieri (2003), Nersessian (1992, 1993, 2007), McMullin (1985), and Mišćević (1992, 2007). This is why it is so important to address Thagard's skepticism: before we accept a cognitive science-based account of how thought experiments justify claims, we must first be sure that the same cognitive science does not also show that thought experiments are not to be trusted. Thagard's position may be

1. I'd like to thank Paul Thagard for his very kind encouragement and discussion in preparing this paper, as well as for doing me the great honor of responding to it.

summarized using his own words: “the made-up thought experiments favored by many philosophers are not evidence at all” (2010a, p. 209). Rather, “philosophical attempts to establish truths by a priori reasoning, thought experiments, or conceptual analysis have been no more successful than faith-based thinking has been. All these methods serve merely to reinforce existing prejudices” (2010a, p. 41).

This paper provides reasons to deny Thagard’s skepticism. I will begin with an outline of his phenomenology of thought experiments, and then address his argument that thought experiments are dangerous and misleading when they are used as evidence. While I reject his argument, I accept the cognitive scientific results on which he draws. In fact, I will show that these same results can be used to make a positive contribution to the field. This paper is therefore an effort to convert Thagard and any who sympathize with his position.

I. Thagard’s Phenomenology of Thought Experiments

Thought experiments for Thagard are “mental constructions of an imaginary situation in the absence of attempts to make observations of the world” (2010a, p. 254). This independence from observation makes their output a priori. Thagard defines a priori knowledge as “Knowledge that is gained by reason alone, independently of sensory experience” (2010a, p. 251).

Thagard argues that a priori methods in general, and thought experiments in particular, function by eliciting “innate ideas, concepts that we have at birth” (2010a, p. 36). This is because thought experiments—especially those in philosophy—work by eliciting intuitions about a concept. What that concept is and what we can conclude about it by this method come only from reflection, not careful scientific investigation via experiment. If we occasionally privilege thought experiments over physical ones, it is because we believe there are some concepts that can be investigated only by non-experimental means. The only concepts thus accessible are innate ones. When we ask ourselves what we would say about people who could split like amoebas (as in Parfit 1987), we are performing conceptual analysis on some pre-theoretical concept of personal identity. If mere reflection yields knowledge about such a concept in a way that does not compete with scientific investigation, that concept must have been innate. Such concepts have been at the center of philosophical debates for centuries, but due to the advent of philosophical naturalism and the growing applicability of cognitive science to philosophical questions (Thagard 2010a, p. 6), it is time to leave these concepts behind for more empirically grounded ones.

Why then, have thought experiments persisted in philosophy and science? Because many philosophers are still blinded by the promise of *neces-*

sary truth. While science only offers contingencies, thought experiments explore the nature of concepts like knowledge, representation, reference, identity, and so on, and they promise that what is discovered will be true without the possibility of empirical falsification.

According to Thagard, this promise is precisely what makes thought experiments so misleading. If the goal of thought experiments is to find necessary truths, then they will not serve their purpose, since there are no necessary truths. He claims, “the notion of necessary truths is just as empty as the notion of the a priori” (2010a, p. 38). Believing the opposite will lead only to wasted time.

But surely not all thought experiments function this way; many are employed by well-known scientists and can be found in science textbooks. For a naturalist like Thagard, this type of responsible thought experiment must be differentiated from the other. This is done by claiming that no thought experiment we find in a science textbook is meant to *justify* a scientific claim. Thagard knows of “no case in science where a theory was adopted merely on the basis of thought experiments” (2010a, p. 39). Speaking in terms I think Thagard might agree with, thought experiments in science only play a role in the context of discovery, and never in the context of justification. Thought experiments might be useful for “suggesting and clarifying hypotheses” (2010a, p. 60) and for “revealing inconsistencies in opposing views” (2010a, p. 39). But under no circumstance should we try to use them to justify the acceptance of beliefs (2010a, p. 60).

Thagard’s attack does not stop here, perhaps because he realizes that many thought experiments focus on abstract but still empirically relevant concepts. After thought experiments that investigate innate concepts and those that are used responsibly by scientists to “pump intuitions” (Dennett 1984, p. 2013), these form a third class of thought experiments that Thagard must address. While thought experiments that focus on innate concepts do little more than waste the time of philosophers, thought experiments meant to justify empirical claims can lead to empirically false and therefore dangerous conclusions.

I should note that we must make this distinction between kinds of thought experiments for Thagard, otherwise we come to a contradiction. On one hand, he claims that a priori propositions are dangerous because they cannot be confirmed or disconfirmed by experience. They are totally disconnected from the world. On the other hand, he claims that a priori propositions are not to be trusted because they are often false, and that this has been shown empirically. He cites our intuitions about Euclidean geometry as an example (2010a, pp. 36–37). These two claims cannot both be true, unless they refer to mutually exclusive sets of thought experi-

ments. Therefore I will continue to assume that Thagard is attacking different sets of thought experiments instead of different features of the same thought experiments. This does not affect his main argument; which in his *The Brain and the Meaning of Life* (2010a) is principally one of analogy between thought experiments and religious reasoning.

Believers of all sorts use religious reasoning such as prayer and meditation in an attempt to find truth about non-empirical religious matters, and this corresponds to the use of thought experiments as a means of probing innate ideas. But they may also try to use such reasoning to investigate empirical matters, which corresponds to the other improper use of thought experiments. These are essentially the two problems with faith that Thagard identifies in thought experiments. In his words, here is the first.

A priori reasoning . . . has the same arbitrary nature as faith. Just as what people adopt as their particular brands of religious faith depends largely on their upbringings and associates, so what people take to be true a priori—what they can imagine—depends on what they have already learned. (2010a, p. 37)

What someone finds intuitive is a function of his or her life experiences. What a mathematician finds obvious about identity or love will be quite different from what a plumber or a priest finds obvious. Since our intuitions are partially the result of chance events that helped to form our systems of beliefs, there can be no objective method of adjudication when we disagree about what is intuitive. We cannot say that someone who has a different intuition about an imaginary case is wrong.

If we focus on the evaluation of a priori methods, we see that:

Without . . . evidential evaluation, use of thought experiments becomes merely the trumpeting of one philosopher's intuitions over another's, a process no more conducive to [empirical] truth than the professions of faith by advocates of rival religious sects. For every thought experiment there is an equal and opposite thought experiment. (2010a, p. 39)

That is, thought experiments are more like subjective reports of opinion than a means of access to the world. The imaginary scenarios we cook up do nothing more than provide rhetorical banners behind which we pronounce our beliefs. The banners themselves justify nothing.

The second problem applies to the empirical use of a priori reasoning, and it was alluded to above: faith and thought experiments lead too often to false beliefs, and so they are not empirically reliable. The strength of this criticism depends on case studies, and Thagard provides many exam-

ples of thought experiments, like John Searle's, that he believes have led science astray (see Thagard, [this issue]).

I think this is enough to show that Thagard sees two different sorts of thought experiments, each with different defects, and not two problems that apply to all thought experiments. Having drawn this distinction, I will leave Thagard's comparison of thought experiments to religion here. His points about thought experiments could be correct independently of whether they also apply to religious beliefs. The real issue is whether the intuitions that play a role in thought experiments can be reliably grounded, and whether the method of thought experimentation is one that is capable of justifying claims, especially in science, where perhaps the greatest danger lies.

II. Against Thagard's Phenomenology of Thought Experiments

Before replying to Thagard's skepticism, there is an important caveat to be made. His points about the evidential power of thought experiments are meant to be limited to philosophical and not scientific thought experiments, because the former are used to justify claims, while the latter are properly limited to the generation of new hypotheses, or finding inconsistencies. However, Thagard's naturalism implies that the methods of philosophy and the methods of science are (or should be) one and the same. Presumably then, Thagard would allow that thought experiments could be used responsibly in philosophy as well as science, to point out contradictions and inspire new ideas. The real issue is that some thought experiments are believed to yield important evidence for claims about the empirical world, and for Thagard this promise is never borne out. Therefore in the remainder of the paper I will not differentiate between thought experiments in science and philosophy, but rather focus on whether or not thought experiments can provide evidence.

In this section I respond to Thagard's notion of a priori reasoning and why he thinks it should not be admitted into science, and then to his understanding of thought experiments.

Thought Experiments as Independent of Experience

As we saw above, thought experiments are a priori for Thagard. By "a priori" Thagard simply means "independent of experience." But there are several ways a proposition can depend on experience. Here are three. 1) A proposition might depend on experience in the sense that some experience is necessary in order to *understand* the proposition. 2) A proposition might depend on experience in the sense that some experience is necessary to *cause* us to entertain that proposition. And 3), a proposition might depend

on experience in the sense that some experience is necessary in order to *justify* that proposition (see Boghossian and Peacocke 2000).

Of interest in the present context is the third sense of dependence. That is to say, a proposition can depend on experience in the first two senses, and still be a priori. Here is how. Consider these two sentences: “The green car is green” and “the green car is *mine*.” The first can be known a priori, even though we would not be able to *understand* this statement without some general experience, for example, with colors. This is the first sense of dependence. The second sentence is different. While it does rely on previous experience for comprehension, it also calls for some specific experience to justify our knowledge of it (perhaps the memory of signing ownership papers). This is dependence in the third and relevant sense.

Let’s look at the second sense of dependence. Suppose we come to know that *modus ponens* is valid after seeing a proof written out in symbols on a piece of paper. Those symbols *cause* us to entertain a proposition, and we need various perceptual capacities for this to happen. This is the second sense of dependence. But perceptual experience is not what justifies the conclusion. That is done by the proof itself (Peacocke 2000, p. 255). This type of knowledge is independent of experience in the third, relevant sense.

Thagard’s conception of the a priori is conceptually underdetermined; he doesn’t distinguish between these senses, and he assumes that dependence in any of them makes a proposition or method a priori. This is what leads him to claim that a priori beliefs must be based on innate ideas. Perhaps he is right that anything independent of experience in all three senses would have to be innate, but there are good reasons to think these senses can and should be separated. Therefore, thought experiments can be linked to experience in important ways, even if Thagard is correct to characterize them as a priori.

However, if we suppose that a priori reasoning really is restricted to innate ideas, we could still perhaps show that they are connected to experience in an appropriate way. If innate belief-structures are the product of evolution, they may well have originated from human experience, though indirectly through the operation of natural selection on the belief structures of generations of our ancestors (see Sorenson 1992). This means that we can also deny Thagard’s inference by denying the second step: innate ideas do not obviously correspond to Thagard’s notion of independence from experience. Even if thought experiments are a priori and function by the use of innate ideas, they need not be independent of experience in a way that disconnects them from reality.

I conclude that Thagard’s characterization of the a priori is too crude to be of any help in assessing the evidential significance of thought experi-

ments. “Independent of experience” is a criterion that can be satisfied without the need for innate ideas. And once we see how much empirical content can be involved in a priori reasoning, it begins to look more plausible that such reasoning could stand as evidence for claims about the world.

Thought Experiments as a Means to Necessary Truth

There are at least two problems with the idea that thought experiments are meant to lead to necessary truth. First, it imputes a specific aim or goal into the heads of thought experimenters everywhere. This is problematic. Second, it assumes that some of the philosophers who discuss thought experiments believe that this is the method’s aim, and I think this is false.

There are figures in the history of philosophy who might have agreed with Thagard that thought experiments can reveal necessary truths (perhaps Kant is an example). But the majority of contemporary philosophers do not see this as one of the primary functions of such devices. I will begin with John D. Norton, who argues that thought experiments are nothing but deductive, inductive or informal arguments. While some deductive arguments do attempt to yield necessary truths, including perhaps mathematical arguments, it is certainly false to say that inductive and informal arguments do as well. And Norton claims that many thought experiments are really these sorts of argument, which provide only some less-than-certain degree of confirmation. Presumably, Thagard will not object to thought experiments if they are arguments.

Besides Norton, there are philosophers who believe that thought experiments are manipulations of mental models. The claim here is that thought experiments create a

. . . representation of the problem, i.e., [a] representation giving the opportunity for the subject to manipulate the problem situation (in his head of course) in a particularly easy fashion, and so that it makes it easy to mobilize [the] subjects’ cognitive resources—skills, implicit background knowledge, perceptual beliefs, etc., in a way superior to regimented reasoning. (Mišćević 1992, p. 224)

In other words, thought experiments are actions that connect and manipulate representations in a way that brings us to something new. Their output is only as reliable as the representations they manipulate, and the cognitive mechanisms they use. Since our representations and cognitive mechanisms are imperfect, we do not expect necessary truth from thought experiments thus conceived. These philosophers are, broadly speaking, naturalists like Thagard, and I will argue below that Thagard should ally himself with their cause.

But there are also rationalists in the debate, and perhaps Thagard is concerned with these philosophers, who claim that the operation of reason alone can produce new knowledge. There is even a Platonist who argues that thought experiments can provide a glimpse into Plato's heaven. In the past, there have been philosophers who believed that whatever is revealed or deduced by the power of reason is necessarily certain. But Thagard's opponent is not Plato, Descartes, or Kant. We should note that almost all modern day rationalists accept fallibilism about a priori knowledge (see, e.g., Bealer 2002; Bonjour 1998; Friedman 2001; Peacocke 2000). This includes Platonists and non-Platonists who believe that thought experiments are a priori (see Arthur 1999; Brown 1991).

Having quickly surveyed the broad position types found in the literature, it is not clear that there is anyone who holds the thesis that thought experiments are meant to reveal necessary truths. If no one holds this thesis, then Thagard is attacking a straw person. Even still, there is something to be said for the strawperson's position.

How might thought experiments lead to necessary truth in an epistemically responsible way? One way is for thought experiments to tell us something about the essence of a concept. When Thagard discusses thought experiments as a means of doing conceptual analysis, we are given examples that fit this characterization, i.e., we are given thought experiments that seek necessary and sufficient definitions for our concepts. If the concept in question is something like "electron," then I believe Thagard is right to be skeptical. We should not pretend that mere thought could reveal the necessary or sufficient conditions that would inform us about the members of this concept's extension, or the characteristics of those members. However, it is another matter insofar as we deal with concepts that are more or less of our own construction. There are concepts that are more defined than discovered, and the necessity that results from a thought experiment about these concepts can be *conceptual* rather than metaphysical. It is wrong to ignore this distinction and claim that all thought experiments that function by conceptual analysis aim at necessary truth and are epistemically irresponsible for this reason. This is because conceptual necessity is not dangerous: it is often nothing more than the establishment of a useful tautology. Perhaps Thagard would claim that those thought experiments that deal only in conceptual necessity are permissible. He might say that these are the ones that help in revising our theories and developing hypotheses. But if that is the case, why denounce thought experiments for their reliance on or kinship with conceptual analysis, as Thagard does?

Perhaps Thagard would challenge conceptual thought experiments by denying the distinction between conceptual and metaphysical possibility

or between conceptual and empirical concepts. In this case, he must provide some account that explains how formal concepts, including those of mathematics and logic, can be reduced to empirical ones. This is an old challenge for empiricists to which Thagard proposes no answer.

This is not to say conceptual analysis is without its problems. I share Thagard's skepticism concerning instances in which necessary truths about the physical world are purportedly *established* by nothing more than mere thought. But such cases are the exception: it is rare to find a thought experiment that claims to have established a fact about the physical world (or an empirical concept) without relying on supplementary evidence at all. A priori reasoning provides varying degrees of support; from suggestion to establishment. And we should not make the mistake of thinking that because a thought experiment deals with something that is necessary if it is true (e.g., x and y are related by the physical law of nature z), it is itself meant to be the sole ground of that necessity. Evidence for a claim that is thought to be metaphysically necessary can still be weak; it need not attempt to establish the necessity on its own in an intellectually irresponsible way. Einstein cannot run as fast as a wave of light in his mind because this would force the wave to become static and no longer what it is, a moving wave (Einstein 1949). This thought experiment can be taken as evidence that there is something special about electromagnetic waves that limits our ability to change their apparent motion by changing our own (see Norton 2012). But while Einstein's thought experiment might provide only *prima facie* evidence for this claim, this is still evidence. The same can be said about a priori reasoning in philosophy, for example, concerning Thomson's famous violinist thought experiment that supports a woman's right to abortion. If we are brought to believe that we can remove a person who has been connected to us without consent in order to save that person's life, this is *prima facie* evidence that the right to life is separate from the right to bodily integrity. That is, abortion is in some cases primarily an exercise of the right to bodily integrity, and only secondarily a displacement of the right to life (Thomson 1971). Thus, a priori reasoning can be used to make a case for a metaphysically necessary proposition, but there is no reason to claim that it must establish such a necessity all on its own.

This section has addressed some of Thagard's arguments against the first class of thought experiments—those that are criticized for being isolated from experience and drawing only on subjective opinions about innate ideas to yield necessary truth. We were able to conclude that thought experiments aren't isolated from experience, even if they are a priori. And while they need not operate solely using innate ideas, it may not be a

problem even if they do. Finally, they do not always aim at necessary truth, although even when they do, they are not always objectionable.

Against Thagard’s Definition of Thought Experiments

Perhaps Thagard’s skepticism about thought experiments is guided by a poor definition of thought experiments. He defines thought experiments as “mental constructions of an imaginary situation in the absence of attempts to make observations of the world” (2010a, p. 254). For one, this definition ignores the important performative aspect of a thought experiment. In many cases we do not merely construct an imaginary situation, we *do* something with it.

Second, Thagard’s definition implies that all thought experiments must remain physically unperformed for them to count as thought experiments. But surely if we physically perform Galileo’s thought experiment about falling bodies, this does not make Galileo’s thought experiment stop being a thought experiment. And Thagard himself notes that this “thought experiment” has been physically performed (2010a, pp. 25–38), so I believe that Thagard implicitly realizes that later attempts to make an observation is not especially relevant for what makes something a thought experiment.

Finally, this definition hardly suffices to distinguish thought experiments from any kind of imagined scenario or work of fiction. Whatever the relation between fiction and thought experiments turns out to be, it will do no good to claim that every imagined scenario is a thought experiment. This is because such a broad definition misses what is really interesting about thought experiments, namely, that they aim to extend our knowledge. Staring up at the clouds and imagining creatures locked in mortal combat must be distinguished from the practice of thought experimentation, for practical epistemic purposes. The Einstein-Podolsky-Rosen thought experiment is not in the same business as cloud-gazing.

To summarize, Thagard’s definition of thought experiments will not do. I urge him to adopt one of the more standard definitions in the literature. This would not substantially affect his arguments against thought experiments, because what he objects to isn’t the mere formation of mental constructions, but placing evidential weight on the intuitions elicited by such constructions. Let us then see what can be said against his specific objections to the epistemic use of thought experiments.

The Method of Thought Experiments

I made the case above that thought experiments do not usually aim to establish necessary truths. So what else might they do, and how? Brown suggests at least the following roles for thought experiments: they may *test*

1 LINE SHORT
REGULAR

scientific conjectures (by refuting or confirming them); *illustrate* theories; *simulate* or *uncover* natural phenomena, and *create* new phenomena entirely (1991). Roy Sorenson claims that thought experiments are mainly used to test the modal status of propositions, and “the favorite use of thought experiment is to establish a possibility” (1992, p. 36). Sorenson adds that thought experiments may also be used to control extraneous variables, invert ideals of natural order, and serve as a “master test for conceptual analysis” (1992, pp. 11–16). John Norton sees thought experiments as arguments which expand our knowledge by induction and deduction (2004). Catherine Elgin claims that they “exemplify” properties and relations in a way that makes interesting or relevant features of the world stand out [this issue].

Thagard disagrees. He sees only two main *modus operandi* for thought experiments, corresponding to the two improper uses of thought experiments that I distinguished at the beginning: the eliciting and broadcasting of arbitrary intuitions, and the inference from what we can imagine to what the world must be like. We have already discussed the first, so I will now turn to the second. My response is that not all thought experiments use this inference, and those that do are not necessarily unreliable for this reason.

From Conceivability to Possibility

In Galileo’s falling bodies thought experiment, we do not ask ourselves whether the situation presented is conceivable or possible, because we could easily make it *actual*. Its conceivability might be a necessary precondition of our performing the thought experiment, but on its own, it is irrelevant for the justification of the thought experiment’s conclusion. One thing that helps to justify the thought experiment is that it predicts what would actually happen if it were performed, ignoring considerations like air resistance (see Buzzoni 2008). Other things that justify the thought experiment include having empirically well-supported premises, there being a relationship between those premises and the thought experiment’s conclusion that would be approved by logicians (Norton 2004), and if all its assumptions are acceptable to its opponents (Popper 1959).

A more complex example is Schrödinger’s Cat (Schrödinger 1935). We might think that the point of the thought experiment is to see that it is possible for a cat to be neither dead nor alive nor neither nor both, given that we can conceive of it as connected to the type of quantum system Schrödinger describes. But in fact, we cannot properly conceive of the cat in superposition, just as we cannot conceive of a particle in superposition. Equally, it doesn’t really matter if such a set-up is possible. Again like Galileo’s, it looks like it is actual (Romero-Isart et al. 2010). The point of the

thought experiment is rather to show that when we combine the Copenhagen interpretation, which takes the superposition of an object to be the occupation of all possible states of the observable variable (e.g., alive, dead, both, neither) until observed, with the Schrödinger equation, which implies that any two systems may be “entangled,” then we are forced to admit that macroscopic systems (like cats) must be capable of “resolving” in the same way that wave packets do. This is not an exercise in innate ideas. It does not point out a contradiction or generate a new hypothesis, and it does not lead us astray by concluding that something is possible because we can imagine it. Instead, the thought experiment seems to rely on perfectly standard types of inferences. It forces physicists into taking a stand on exactly *when* we should say that the cat has survived or perished, and what role observers play in the “collapse” of entangled quantum systems. If the cat only “resolves” when an observer opens the box, then what does the cat, itself an observer, see?

There are many other thought experiments whose justification does not rely on the conceivability-possibility inference (e.g., Poincaré’s Disk, Einstein’s Elevator, EPR, etc.), but I would like now to address those that do. In the last two decades, much work has been done on the inference from conceivability to possibility, and it is generally agreed that the former is at most only a *guide* to the latter. The relation is not one of entailment. Rather, conceiving imaginary scenarios engages our modal faculty and enables fruitful modal discourse (see e.g., Gendler and Hawthorne 2002; Yablo 1993). Thought experiments that use this inference should therefore be understood as attempting to provide only a weak form of evidence for a modal claim, although evidence nonetheless.

Without trudging too deep into the murky waters of modality, it seems plausible that we have at least some fallible way of knowing what is possible. This ability, however it works, is responsible for our capacity to reason counterfactually, which seems to be a necessary precondition of scientific thought in general (Buzzoni 2008). If we couldn’t imagine different ways the world *might* be, we couldn’t find out by experiment what it was really like. For instance, we could not envisage instruments for testing possible values of a variable, or experimental setups that make such measurements possible. As long as conceivability is used as an aid to this faculty, and as long as we assume that science’s methods are the *right* methods (which Thagard does), it should not be labeled dangerous or misleading.

Up to this point I have tried to defend a priori reasoning and thought experiments against Thagard’s skepticism. Now I will show that these tools of reasoning are reliable and perhaps even necessary for doing good science, according to Thagard’s own conception of “good” science.

Thought Experiments as Weakly Evidential

For Thagard, the archetype of good reasoning is scientific reasoning. He claims that we can delimit good from bad science as follows:

In general, we can use descriptive information to help generate normative conclusions whenever we can identify the appropriate goals. If the appropriate goals of science are truth, explanation, and prediction; and if the history of science reveals that experiments and inference to the best explanation are the best practices for achieving these goals; then these practices are normatively justified as what scientists ought to do. (2010a, p. 175)

Thagard argues that inference to the best explanation and experiment are the practices that achieve the goals of science, so they are normatively justified in science and presumably any other field with similar goals. But then the practices that make inference to the best explanation and experiment possible would also (indirectly) lead to truth, explanation and prediction. Is there anything that makes inference to the best explanation possible? Thagard claims that we find the best scientific explanation by seeking coherence and avoiding inconsistency (2010a, pp. 21–22). And these ancillary goals, Thagard admits, can be achieved using thought experiments. He says, “thought experiments are fine for suggesting and clarifying hypotheses” (2010a, p. 60), and also for “revealing inconsistencies in opposing views” (2010a, p. 39). Thus, thought experiments can (at least indirectly) help us to reach truth, explain phenomena, and make predictions. Insofar as thought experiments do this, they provide evidence for claims, explanations, or predictions. This shows that Thagard is wrong to claim that thought experiments can or should play no evidential role in science. According to his own definition of good science, and his own characterization of the roles that thought experiments can play, thought experiments are a useful evidential tool for the scientist.

Thought Experiments as Strongly Evidential

If there is an *experimental* side to thought experiments then they are evidential, since Thagard’s definition of “evidence” highlights information gained by experiment (see below). I think I can show that there is such a side to thought experiments, first by expanding Thagard’s definition of experiment, and then tightening it. Thagard defines experiments as “planned manipulations” that alter only a few features of a system to “be able to identify causes and effects.” They are “repeatable,” and they make possible precise quantitative measurements (2010a, p. 25). He is right that thought experiments are not experimental under this definition. For one, they do not usually make precise quantitative measurements possible.

However, I would point out that many accepted physical experiments also lack this quality—e.g., the celebrated experiments conducted by Pasteur that have been taken to disprove the thesis of spontaneous generation (see Conant 1953). These experiments showed that life will not spring from inanimate material such as boiled meat if we remove the presence of air containing “spores” or “germs.” Fleas, and maggots, and mice come from other fleas, maggots and mice, not from dust or dead meat or dirty rags. Under the condition that all experiments must make possible precise quantitative measurements, experiments like those of Pasteur are disqualified.

Leaving this criterion out of Thagard’s characterization, we must still face the others, and it is true that many thought experiments do not identify causes or effects. However, once again, many physical experiments fare no better—e.g., Michelson and Morley’s famous experiment (1887) suggested only that the ether theory was incorrect; it suggested no cause or effect. Examples of this kind can be multiplied by considering other important experiments with negative conclusions.

We are left with repeatability, which thought experiments must be able to satisfy, since otherwise we cannot make sense of certain historical episodes like the Bohr-Einstein debate over the Clock-in-the-Box (Bishop 1999). In this case, Einstein and Bohr certainly disagree about *something*. And in order for their disagreements to resolve rationally, the disagreement must be about the *same thing*. Therefore when Bohr replies to Einstein’s thought experiment using another similar thought experiment, a naturalist like Thagard is forced to presume that these scientists repeat the same thought experiment with increasing precision and care, otherwise we will have to say that Einstein and anyone else who accepted Bohr’s reply was irrational for doing so.

Could Thagard change his definition of experiment to exclude all thought experiments and retain all physical experiments? I think this would be very difficult. “Experiment” is a general term for good reason. This consideration makes it more and not less likely that thought experiments will count as experimental.

Of course, we should also concern ourselves with the characteristics that *do* unite the practices that fall under the spectrum of applicability of “experiment.” Let us first admit that actions can be more or less experimental. Even still, all experimental activity seems to test the change in some few variables while controlling others. It also always takes place in response to the formulation of a hypothesis, and it relies on some idealization. In a computer simulation, we manipulate an idealized representation of a system with the hope that its output will be reliable with respect to that system. This is at least *somewhat* experimental, according to many

philosophers (see Guala 2002; Morgan 2003, 2005; Lenhard 2007; Parker 2009; Morrison 2009; Giere 2009; Barberousse et al. 2009).

Once we acknowledge the seemingly experimental properties of computer models, we should note that it is a short step from a computer model to a mental model or thought experiment (see Di Paolo et al., 2000). Each of these methods manipulates representations using more or less specific inference patterns, both reflect heavily the theory from which they are developed, and both rely on structural analogies between target and represented system, as opposed to the physical analogies exploited in laboratory experiments. My point in all this is that Thagard must provide a principled way to set apart computer simulations and thought experiments epistemically, or he will be forced either to include both as evidence-producing tools of scientific reasoning, or let go of computer simulations, which have formed a major part of his work since the 1980s. In either case, he cannot disqualify thought experiments as experimental without some losses elsewhere in his philosophy.

Evidence

Finally, what exactly does Thagard mean when he says that thought experiments cannot count as evidence? Popular notions of evidence are quite divergent. Some characterize evidence as whatever provides justification for a proposition (Kim 1988), or the set of all one's knowledge (Williamson 2000), or sense-data (Russell [1912] 1997), or observation statements (Quine 1968), or the set of all one's occurrent thoughts (Conee and Feldman 2004), or physical things like a murder weapon. And there are many others. On several of these characterizations, thought experiments are unproblematically evidential. Thagard's own definition of evidence is: "information gathered by careful observation, especially through scientific experiments" (2010a, p. 252). Can thought experiments produce information gathered by careful observation? That depends on what we count as careful observation. If Thagard means to exclude introspection, then we lose much of our evidence for claims like "I am hungry" or "I am upset." But perhaps we do not wish to count introspection as evidence. Even still, there are many other cases where "observation" seems inappropriate to characterize an instance of evidence. For example, we receive evidence from arguments, computer simulations, mathematics, climate models, and many other methods that do not operate by observation. However these methods function epistemically, whether by manipulating representations within predetermined system constraints or transforming propositional content according to truth-functional rules, they do not function by observation. The output of any argument, simulation, proof, model or thought experiment is trivially observable, if we write it down or have it

presented on a screen, but that is not what justifies its output. *Evidence* is a normative epistemic concept related to justification, and *observation* is too narrow to give us that normativity. Again we find ourselves in a situation analogous to the one above: either we constrain evidence to information gained by careful observation in a sense that excludes thought experiments, and lose the evidential capacity of other means that we normally think of as evidential, or we recognize that the notion of evidence is flexible for a good reason, and this again makes it more and not less likely that thought experiments can serve as evidence.

It has been my aim to show that none of Thagard's arguments against the evidential significance of thought experiments are decisive. In what follows I aim to convert Thagard by considering some of the results from cognitive science on which he draws in order to explore the ability of thought experiments to provide evidence.

III. Cognitive Science, Thought Experiments and Mental Models

With Terrence Stewart, Thagard has presented a new account of human creativity (2011). It is interesting to note that in this article the authors cite Nancy Nersessian's account of mental models approvingly. Nersessian defines a mental model as a "structural, behavioral, or functional analog representation of a real-world or imaginary situation, event or process" (2008, p. 93). Thagard and Stewart go on:

We agree with her contention that many kinds of internal and external representations are used during scientific reasoning, including ones tightly coupled to embodied perceptions and actions . . . our account is certainly compatible with Nersessian's (2008, p. 135) claim that conceptual changes arise from interactive processes of constructing, manipulating, evaluating, and revising models. (2011, p. 25)

This is telling because Nersessian is well-known for identifying thought experiments with mental models (1992, 1993, 2007). Thagard is willing to accept that mental models can produce new knowledge, since it is through their use that creative new conceptual combinations are made possible. To explain how we gain understanding from thought experiments in a way that is consistent with Thagard's work in cognitive science, I will assume the identification of thought experiments with mental models.

This is a legitimate assumption because even opponents of this view do not deny that thought experiments may be portrayed usefully as mental models. They admit that such a characterization might even be the most descriptively accurate. Their objection is only that this characterization

does not do a *better* job of capturing the epistemically interesting features of thought experiments (Hacking 1992; Norton 2004; Brown 2007). In what follows, then, I will assume that thought experiments may be characterized as mental models. To explore this new account, I will first present Thagard's understanding of mental models, and then how we learn from them.

For Thagard, mental models are representations, "consisting of patterns of activation in populations of neurons" (2010a, p. 78; 2010b, p. 447). A neural population is "a collection of neurons that are richly interconnected," and a population of neurons "represents something by its pattern of firing" (2010b, p. 450). Mental models and concepts are often "produced by, and maintain some of the structure of, perceptual inputs" (2010a, p. 78), and presumably, the way they maintain the structure of their objects is what makes them reliable. The idea is that "a pattern of activation in the brain constitutes a representation of something when there is a stable causal correlation between the firing of neurons in a population and the thing that is represented, such as an object or group of objects in the world" (2010b, p. 450). That is, neural populations are causally connected to the things they represent because they preserve structure. If I see something to the left of something else, and I represent that situation to myself, my representation of one object will be in a neural population that is physically (or if you like, neurally) to the left of the other, and this will instantiate my concept "to the left of." The same goes for temporal and other kinds of structures which may be preserved in thought.

We dynamically manipulate these neural structures to create mental models. Mental models are a product of the interaction of higher-level mental representations like "cause," which are themselves products of lower-level representations stemming from the senses or emotional faculties. We know that it is higher-level interactions that produce the mental model because mental models are conscious, and lower-level interactions are not. What Thagard calls "top-down" (mind-to-neuron) processes monitor and guide the way the model is run. These top-down processes work with the bottom-up processes, such as simple representation, and they all run together, creating feedback loops which ensure that what is represented and manipulated maintains its structural similarities to the target system.

Mental models yield knowledge, according to Thagard, by enabling the generation of genuinely new concepts (like *sound wave* and *wireless email*, 2010b, pp. 3–4). What happens is that any number of conceptual, linguistic, perceptual, or emotional representations may be combined in a "kind of twisting together of existing representations." This Thagard calls "convolution" (Thagard and Stewart 2011, p. 2). The mathematical func-

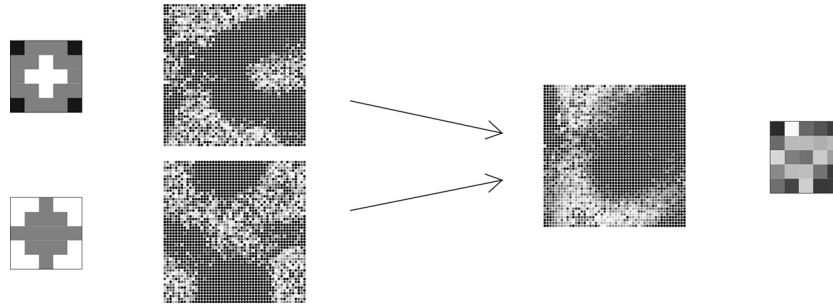


Figure 1. Convolution Occurring In Simulated Neurons” (Thagard and Stewart 2011, 15)

tion from which this process gets its name takes several functions as its domain, and yields a blended overlap as its output. Thagard wishes to characterize concept combination in terms of this function, and to do this he shows that we can make a realistic computer model of human creativity using it. Our representation process begins by converting sensory input (e.g., light and sound waves) into electrical impulses in the brain. His computer simulation uses groups of “nodes” which are designed to mimic the stochastic nature of real neurons. These represent different types of input into the language of vectors by coding it using a series of “natural transformations” (2011, p. 11). He can then make the model “represent” and convolute sensory input, as in Figure 1.

Suppose that a visual system is presented with the two visual stimuli on the left. These will be converted into vectors and represented by a network of nodes. The representations are then convoluted to create the new neural network on the right, which can be translated into a new visual output (far right). Thagard points out that “the convolution of the two input patterns is not simply the sum of those patterns and, therefore, amounts to an emergent binding of them” (2011, p. 15). This process reliably produces new content, which becomes a candidate for knowledge.

Creative convolutions are those that are accompanied by the “AHA!” feeling we get when we suddenly realize something important. This response is an emotional reaction triggered by the experience of coherence between what a thinker wants or needs, and what the new combination makes possible. This emotional response becomes part of the convolution which includes the new concept or idea. Thagard and Stewart sum it up this way: “The AHA! experience is not just a side effect of creative thinking, but rather a central aspect of identifying those convolutions that are potentially creative” (2011, p. 19).

With the rudiments of this account now in place, I want to sketch how it fares in comparison with other cognitive science-based accounts of thought experiments.

Thagard's description of mental modeling can be used to update Nersessian's mental models account of thought experiments. She writes,

When a thought experiment is successful, it can provide novel empirical data. These data are novel in the sense that although they are contained in current representations, the means to access them were not available until someone figured out how to conduct the thought experiment. (2007, p. 127)

An account based on this interpretation of Thagard's work could illuminate the notion of "opening a means of access." A mental model brings us to a new conceptual combination by connecting patterns of activation in different neural populations. The information is already contained in the current representations, and it is "made available" when the patterns are convoluted into something useful. This is brought to our attention when an emotional reaction promotes the new datum. In some sense a house is *in* the assorted building materials when they arrive to a construction site, but its final functional form is something new, something that wasn't there before, and it was made available by a specific process of combination.

Thagard's research also strongly suggests that mental modeling involves a great deal of non-propositional content. It is multi-modal; that is, it involves input from emotions, linguistic concepts, physiological states and representations. Those who adopt the mental models account of thought experiments agree that the mechanisms capable of bringing us new knowledge, and the unarticulated nature of the resources upon which they draw, are often non-propositional (see Gendler 2004; Mišćević 2007; Nersessian 2007). Thagard's multi-modal theory of mental cognition provides additional evidence for this claim.

And Thagard's account takes this idea even further by specifying the nature of the non-propositional elements: they are non-propositional because they are *activities* of neural populations that refer by something like structural isomorphism. Nenad Mišćević states that mental models seem to have a quasi-spatial character. If Thagard is right, the spatial and temporal character is actually full-blown, since the brain retains both relations in a literal sense. Mišćević is also impressed by the speed and ease with which thought experiments bring us to a conclusion (Mišćević 2007). According to Thagard's system, much of what happens in a thought experiment is done unconsciously. By the time the AHA! moment arrives, there

has already been a great deal of unconscious convolution, which helps to explain how a process so fast could also be reliable.

This support for the role of non-propositional content also counts against the position of Norton, who claims that all thought experiments work via line-by-line steps of propositional reasoning. Norton's claim is meant to be descriptively accurate even concerning the psychological mechanisms of thought experiments. If arguments only function upon propositions, Norton is wrong that thought experiments are merely arguments if thought experiments are mental models in Thagard's sense.

Conclusion

Thagard claims that thought experiments are dangerous and misleading when used as evidence. He understands them as an a priori method aimed at necessary truth. They seem to operate in isolation from careful empirical observation and rely on merely subjective intuition, and are therefore nothing like real experiments. I have argued that these claims are either false or based on false assumptions. Thought experiments are an exciting part of the scientific method, and as philosophers of science it is our job to understand them. This should be a goal Thagard shares, and so I closed with a short discussion of just a few of the interesting directions that someone with Thagard's extensive knowledge of the mechanisms of the mind might pursue.

References

- Arthur, Richard. 1999. "On Thought Experiments as a priori Science." *International Studies in the Philosophy of Science* 13: 215–29.
- Barberousse, Anouk, Franceschelli, Sara, and Imbert, Cyrille. 2009. "Computer Simulations as Experiments." *Synthese* 169: 557–74.
- Bealer, George. 2002. "Modal Epistemology and the Rationalist Renaissance." Pp. 71–125 in *Conceivability and Possibility*. Edited by Tamar Szabo Gendler and John Hawthorne. Oxford: Clarendon Press.
- Bishop, Michael. 1998. "An Epistemological Role for Thought Experiments." Pp. 19–33 in *Idealization IX: Idealization in Contemporary Physics*. Edited by Niall Shanks. Amsterdam: Rodopi.
- Bishop, Michael. 1999. "Why Thought Experiments are not Arguments." *Philosophy of Science* 66: 534–41.
- Boghossian, Paul and Christopher Peacocke. 2000. "Introduction." *New Essays on the a priori*. Oxford: Oxford University Press.
- BonJour, Lawrence. 1998. *In Defense of Pure Reason: A Rationalist Account of a priori Justification*. Cambridge: Cambridge University Press.
- Brown, James R. (1991) 2011. *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*. London: Routledge.

- Brown, James R. 2007. "Counter Thought Experiments." *Royal Institute of Philosophy Supplement* 61: 155–77.
- Brown, James R. 2012. *Platonism, Naturalism, and Mathematical Knowledge*. New York and London: Routledge.
- Buzzoni, Marco. 2008. *Thought Experiment in the Natural Sciences*. Würzburg: Königshausen & Neumann.
- Buzzoni, Marco. 2012. "Thought Experiments from a Kantian Point of View." Pp. 90–106 in *Thought Experiments in Philosophy, Science and the Arts*. Edited by Mélanie Frappier, Letitia Meynell, and James R. Brown. London: Routledge.
- Conant, James B. (ed.). 1953. *Pasteur's and Tyndall's Study of Spontaneous Generation*. Cambridge: Harvard University Press.
- Conee, Earl and Richard Feldman. 2004. *Evidentialism*. Oxford: Oxford University Press.
- Cooper, Rachel. 2005. "Thought Experiments." *Metaphilosophy* 36: 328–47.
- Dancy, Jonathan. 1985. "The Role of Imaginary Cases in Ethics." *Pacific Philosophical Quarterly* 66: 141–53.
- Dennett, Daniel C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. Cambridge: MIT Press.
- Dennett, Daniel C. 2013. *Intuition Pumps and Other Tools for Thinking*. New York: W. W. Norton.
- Di Paolo, Ezequiel. A., Jason Noble, Seth and Bullock. 2000. "Simulation Models as Opaque Thought Experiments." Pp. 497–506 in *Proceedings of the Seventh International Conference on Artificial Life*. Edited by Mark A. Bedau, John S. McCaskill, Norman H. Packard, and Steen Rasmussen. Cambridge: MIT Press.
- Einstein, Albert. (1949) 1979. *Autobiographical Notes*. La Salle and Chicago: Open Court.
- Friedman Michael. 2001. *Dynamics of Reason*. Chicago: The University of Chicago Press.
- Gendler, Tamar S. 2004. "Thought Experiments Rethought—and Reperceived." *Philosophy of Science* 71: 1152–63.
- Gendler, Tamar and John Hawthorne (eds.). 2002. *Conceivability and Possibility*. Oxford: Oxford University Press.
- Giere, Ron. 2009. "Is Computer Simulation the Changing Face of Experimentation?" *Philosophical Studies* 143: 59–62.
- Guala, Francesco. 2002. "Models, Simulations and Experiments." Pp. 59–74 in *Model Based Reasoning: Science, Technology, Values*. Edited by Lorenzo Magnani and Nancy Nersessian. New York: Kluwer.
- Hacking, Ian. 1992. "Do Thought Experiments Have a Life of Their Own? Comments on James Brown, Nancy Nersessian, and David

- Gooding." *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Vol. 2: *Symposia and Invited Papers*: 302–8.
- Kim, Jaegwon. 1988. "What is Naturalized Epistemology?" Pp. 381–405 in *Philosophical Perspectives 2, Epistemology*. Edited by James Tomberlin. Atascadero: Ridgeview Publishing Co.
- Lenhard, Johannes. 2007. "Computer Simulation: The Cooperation between Experimenting and Modelling." *Philosophy of Science* 74: 176–94.
- McMullin, E. 1985. "Galilean Idealization." *Studies in History and Philosophy of Science* 16: 247–73.
- Meinong, Alexius. (1907) 1973. "Das Gedankenexperiment." Pp. 273–83 in *Über die Stellung der Gegenstandstheorie im System der Wissenschaften*. Edited by Rudolf Haller and Rudolf Kindinger. Graz-Austria: Akademische Druck und Verlagsanstalt.
- Michelson, Albert A. and Edward W. Morley. 1887. "On the Relative Motion of the Earth and the Luminiferous Ether." *American Journal of Science* 34: 333–45.
- Miščević, Nenad. 1992. "Mental Models and Thought Experiments." *International Studies in the Philosophy of Science* 6: 215–26.
- Miščević, Nenad. 2004. "The Explainability of Intuitions." *Dialectica* 58: 43–70.
- Miščević, Nenad. 2007. "Modeling Intuitions and Thought Experiments." *Croatian Journal of Philosophy* VII: 181–214.
- Morrison, Margaret. 2009. "Computer Simulation: The Changing Face of Experimentation." *Philosophical Studies* 143: 33–57.
- Morgan, Mary. 2003. "Experiments without Material Intervention." Pp. 216–235 in *The Philosophy of Scientific Experimentation*. Edited by Hans Radder. Pittsburgh: University of Pittsburgh Press.
- Morgan, Mary. 2005. "Experiments versus Models: New Phenomena, Inference and Surprise." *Journal of Economic Methodology* 12: 317–29.
- Nersessian, Nancy. 1992. "How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science." Pp. 3–44 in *Cognitive Models of Science*. Edited by Ronald N. Giere. Minneapolis: University of Minnesota Press.
- Nersessian, Nancy. 1993. "In the Theoretician's Laboratory: Thought Experimenting as Mental Modeling." *Proceedings of the Philosophy of Science Association* 2: 291–301.
- Nersessian, Nancy. 2007. "Thought Experiments as Mental Modeling: Empiricism without Logic." *Croatian Journal of Philosophy* VII: 125–61.
- Nersessian, Nancy. 2008. *Creating Scientific Concepts*. Cambridge: MIT Press.
- Norton, John. 2004. "Why Thought Experiments Do Not Transcend Em-

- piricism." Pp. 44–66 in *Contemporary Debates in the Philosophy of Science*. Edited by Christopher Hitchcock. Somerset: Wiley-Blackwell.
- Norton, John. 2012. "Chasing the Light: Einstein's Most Famous Thought Experiment." Pp. 123–40 in *Thought Experiments in Philosophy, Science and the Arts*. Edited by Mélanie Frappier, Letitia Meynell, and James R. Brown. London: Routledge.
- Palmieri, Paolo. 2003. "Mental Models in Galileo's Early Mathematization of Nature." *Studies in History and Philosophy of Science* 34: 229–64.
- Parfit, Derek. 1987. *Reasons and Persons*. Oxford: Clarendon Press.
- Parker, W. 2009. "Does Matter Really Matter: Computer Simulations, Experiments and Materiality." *Synthese* 169: 483–96.
- Peacocke, Christopher. 2000. "Explaining the A Priori: The Programme of Moderate Rationalism." Pp. 255–85 in *New Essays on the A Priori*. Edited by Paul Boghossian and Christopher Peacocke. Oxford: Oxford University Press.
- Popper, Karl. 1959. "On the Use and Misuse of Imaginary Experiments, Especially in Quantum Theory." Pp. 442–56 in *The Logic of Scientific Discovery*. London: Hutchinson.
- Quine, Willard Van Orman. 1968. *Ontological Relativity and Other Essays*. New York: Columbia University Press.
- Romero-Isart, Oriol, Matthew Juan, Romain Quidant, and Ignacio Cirac. 2010. "Toward Quantum Superposition of Living Organisms." *New Journal of Physics*. <http://arxiv.org/abs/0909.1469v3>
- Russell, Bertrand. (1912) 1997. *The Problems of Philosophy*. Oxford: Oxford University Press.
- Schrödinger, Erwin. 1935. "Die gegenwärtige Situation in der Quantenmechanik." *Naturwissenschaften* 23: 807–12; 823–28; 844–49.
- Sorensen, Roy. 1992. *Thought Experiments*. Oxford: Oxford University Press.
- Thagard, Paul. 2008. "How Cognition Meets Emotion: Beliefs, Desires, and Feelings as Neural Activity." Pp. 167–84 in *Epistemology and Emotions*. Edited by Georg Brun, Ulvi Doguoglu, and Dominique Kuenzle. Burlington, VT: Ashgate.
- Thagard, Paul. 2010a. *The Brain and the Meaning of Life*. Princeton: Princeton University Press.
- Thagard, Paul. 2010b. "How Brains Make Mental Models." Pp. 447–61 in *Model-Based Reasoning in Science and Technology. Abduction, Logic, and Computational Discovery*. Edited by Lorenzo Magnani, Walter Carnielli, and Claudio Pizzi. Berlin: Springer.
- Thagard, Paul and Terence Stewart. 2011. "The Aha! Experience: Creativity through Emergent Binding in Neural Networks." *Cognitive Science* 35: 1–33.

- Thomson, Judith Jarvis. 1971. "A Defense of Abortion." *Philosophy & Public Affairs* 1: 47–66.
- Wilkes, Kathleen. 1988. *Real People: Personal Identity without Thought Experiments*. Oxford: Oxford University Press.
- Williamson, Timothy. 2000. *Knowledge and Its Limits*. Oxford: Oxford University Press.
- Yablo, Stephen. 1993. "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research* 53: 1–42.